Exploratory Data Analysis

Lecture 7

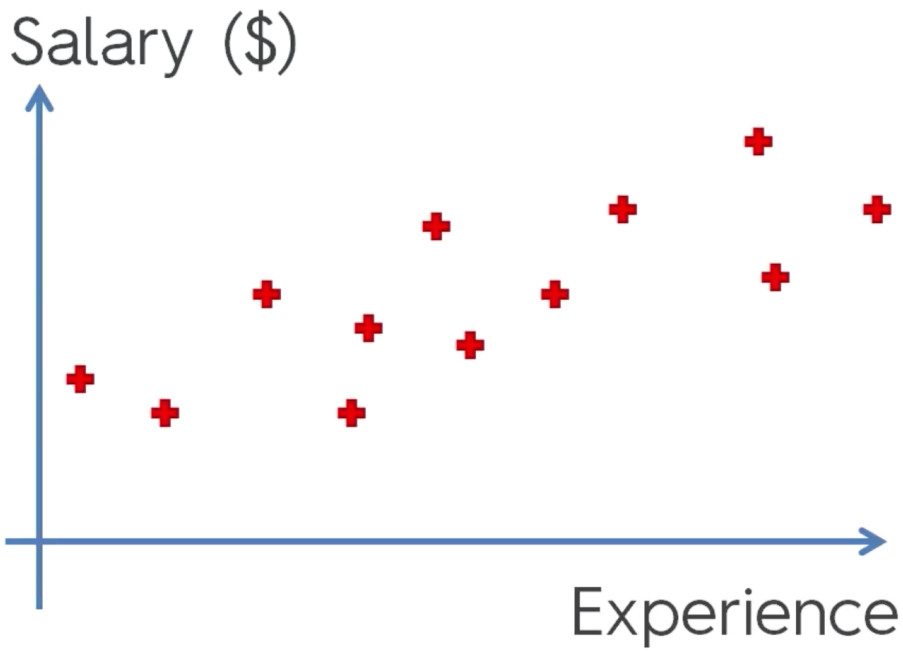
Corina Besliu

Technical University of Moldova

September 21, 2021

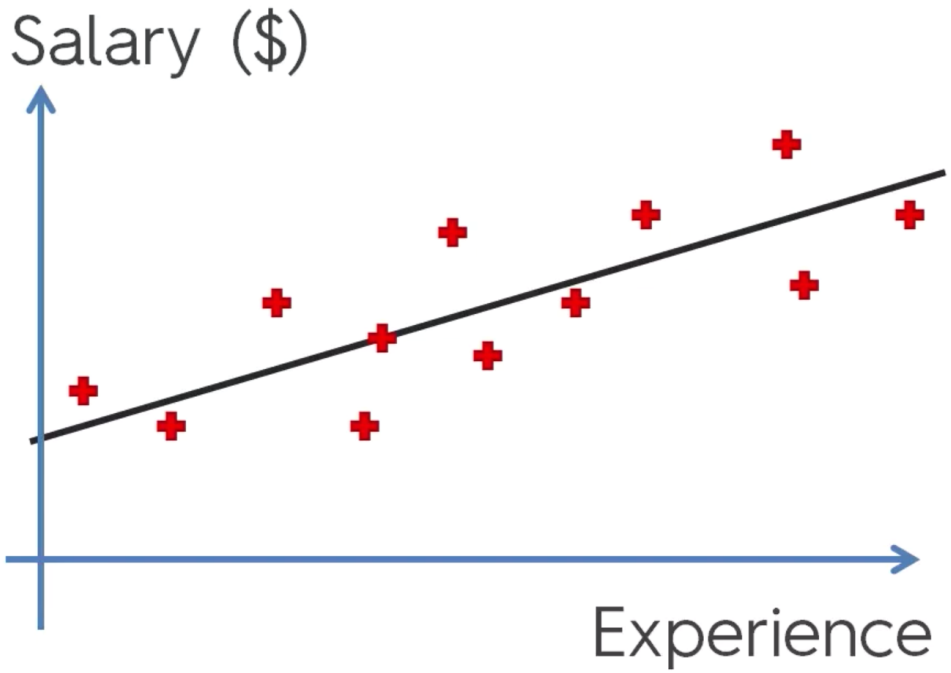
Salary and Work Experience

Simple Linear Regression:



Salary and Work Experience

Simple Linear Regression:

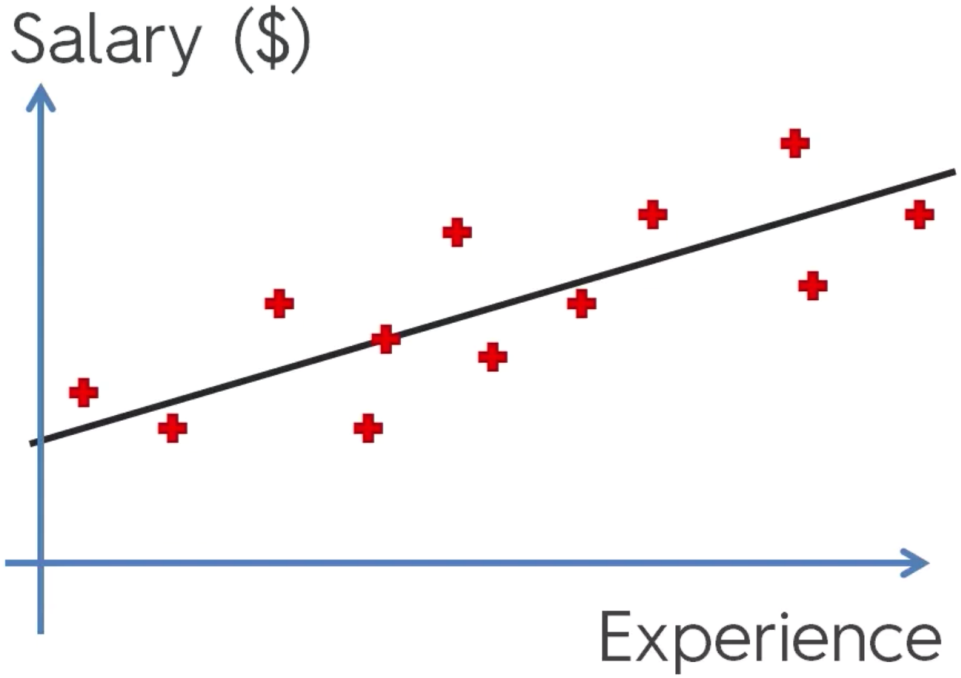


$$y = b_0 + b_1 * x$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

Salary and Work Experience

Simple Linear Regression:

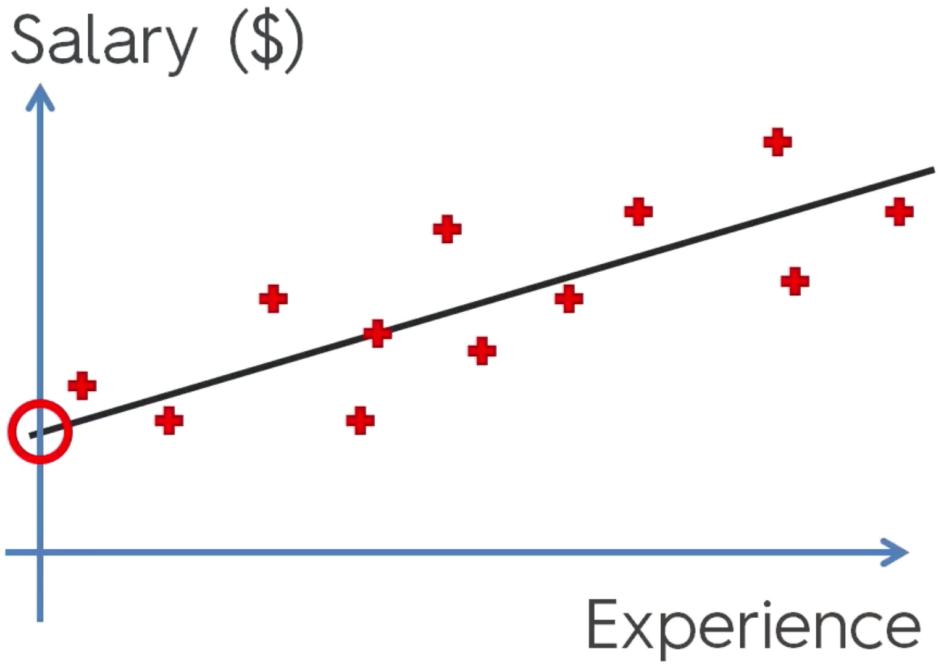


$$y = b_0 + b_1 * x$$

Salary = b_0 + b_1 *Experience

Salary and Work Experience

Simple Linear Regression:

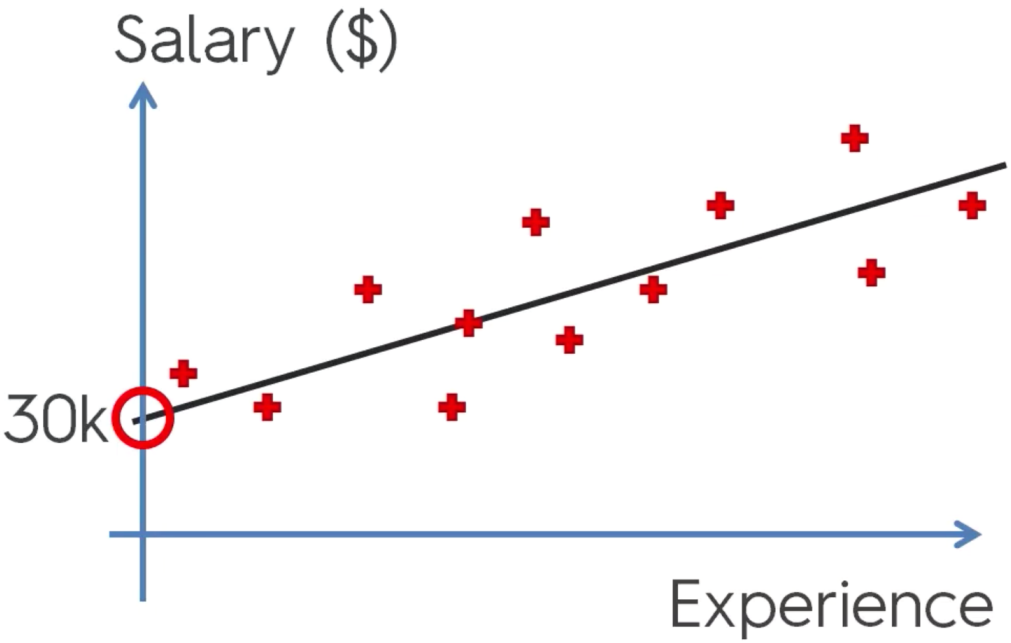


$$y = b_0 + b_1 * x$$

Salary = b_0 + b_1 *Experience

Salary and Work Experience

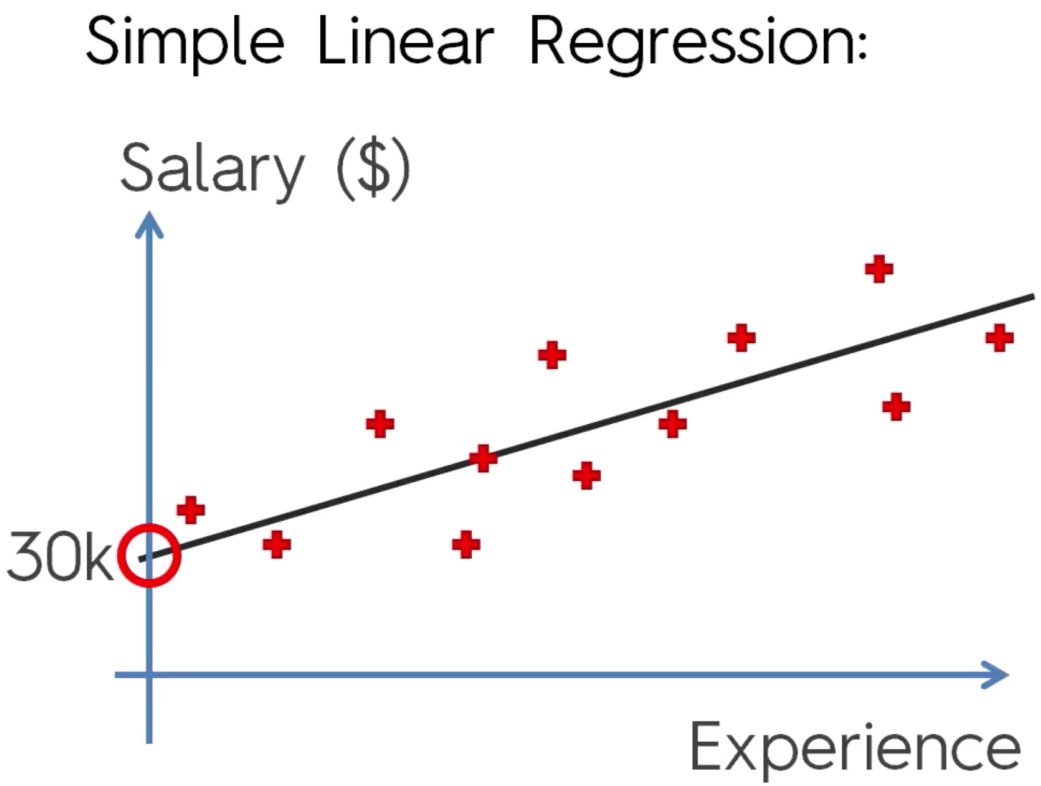
Simple Linear Regression:



$$y = b_0 + b_1 * x$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

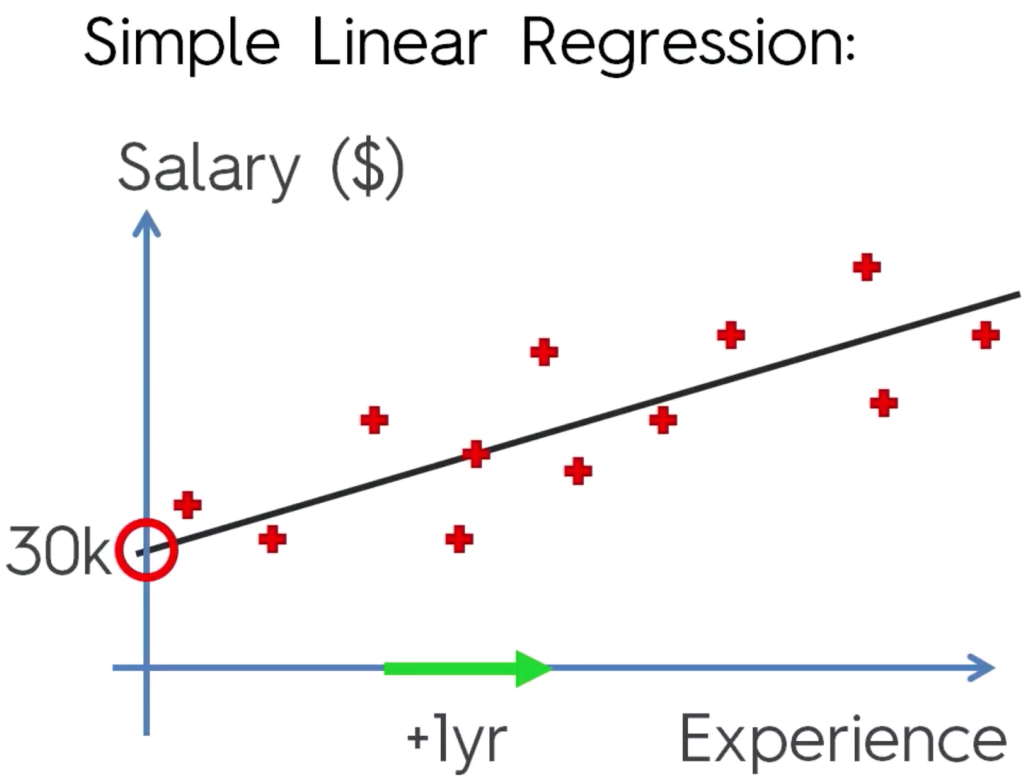
Salary and Work Experience



$$y = b_0 + b_1 * x$$

Salary = b_0 + b_1 *Experience

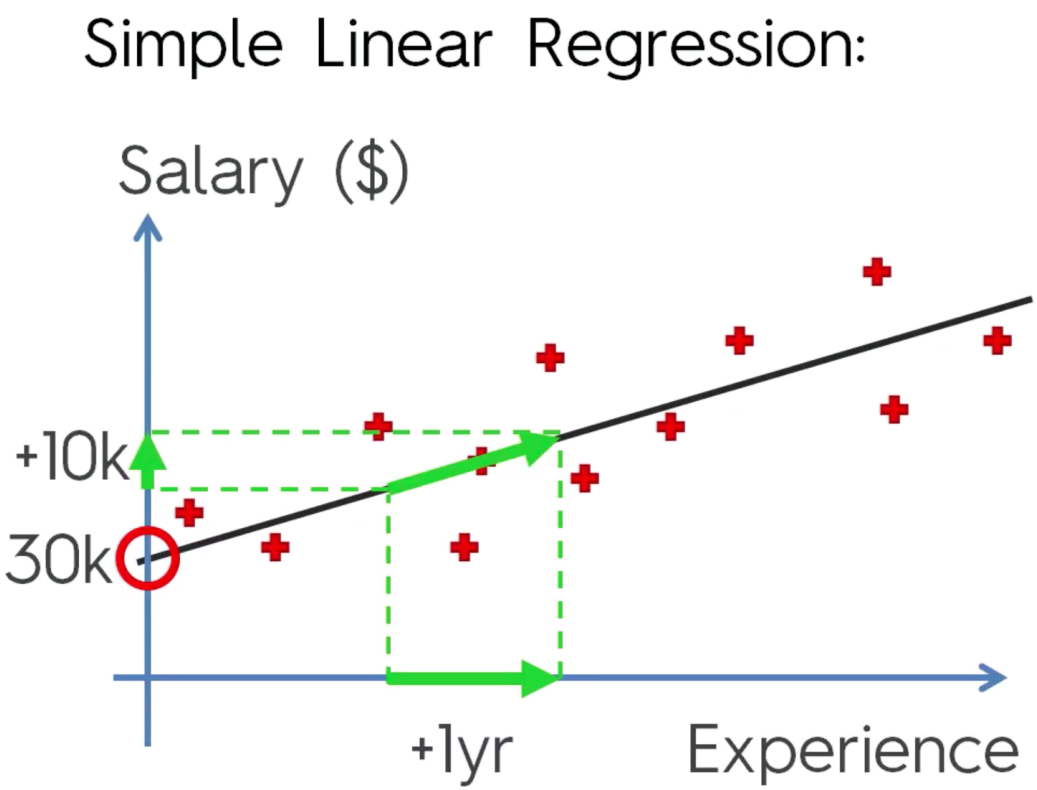
Salary and Work Experience



$$y = b_0 + b_1 * x$$

Salary = b_0 + b_1 *Experience

Salary and Work Experience

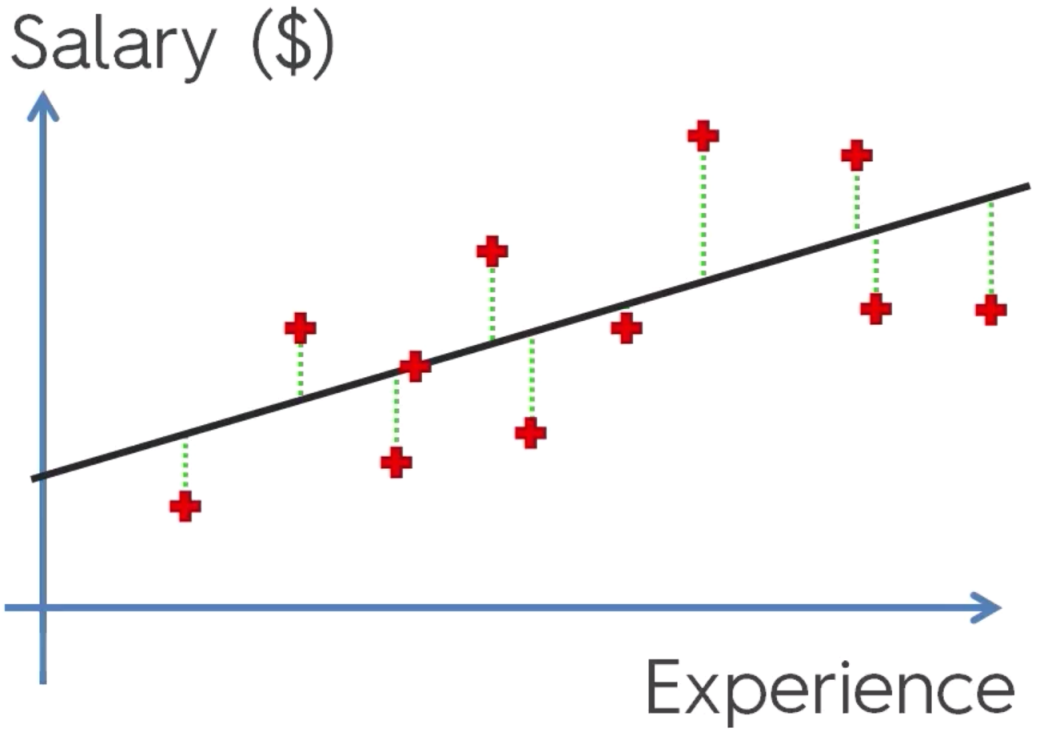


$$y = b_0 + b_1 * x$$

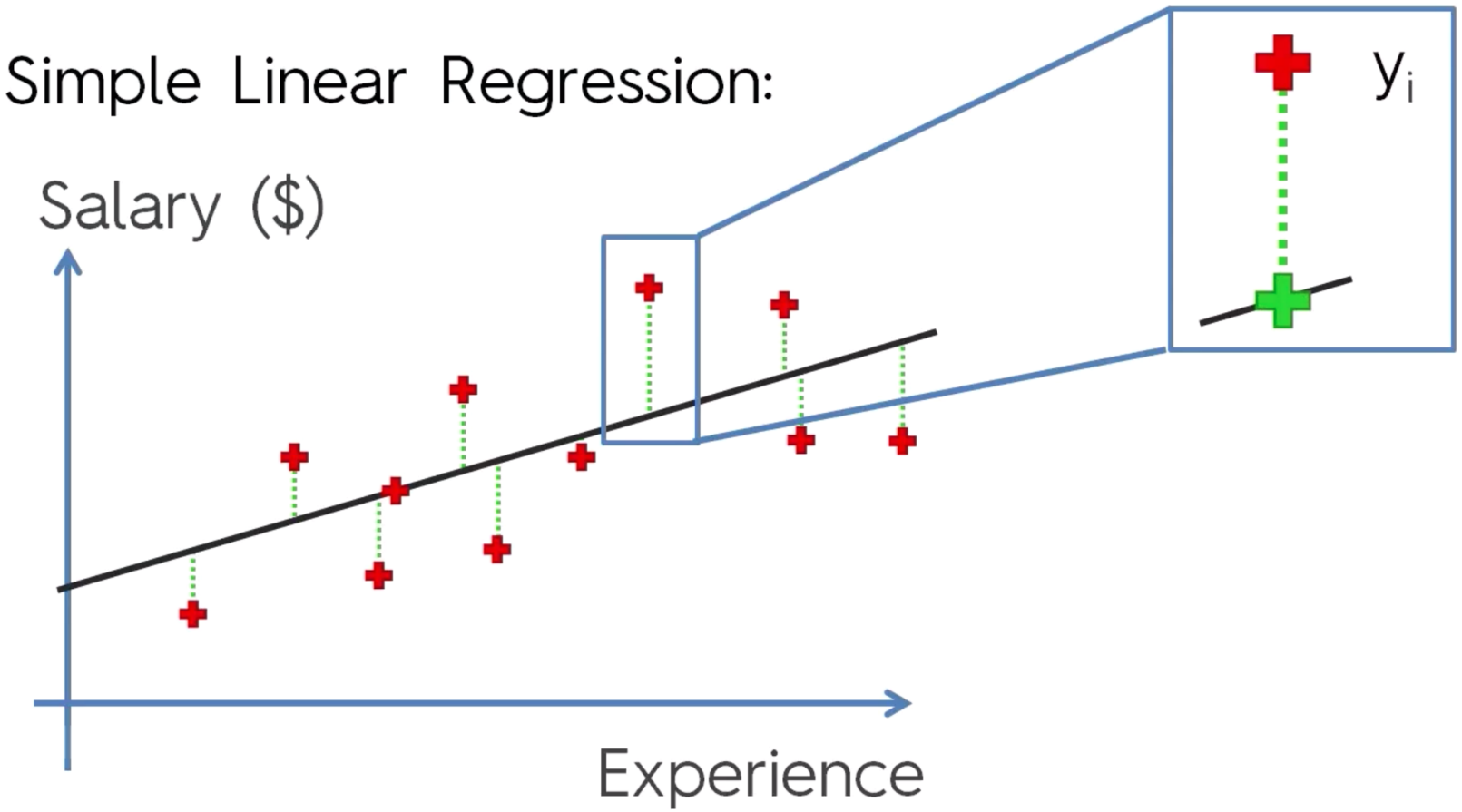
Salary = b_0 + b_1 *Experience

Salary and Work Experience

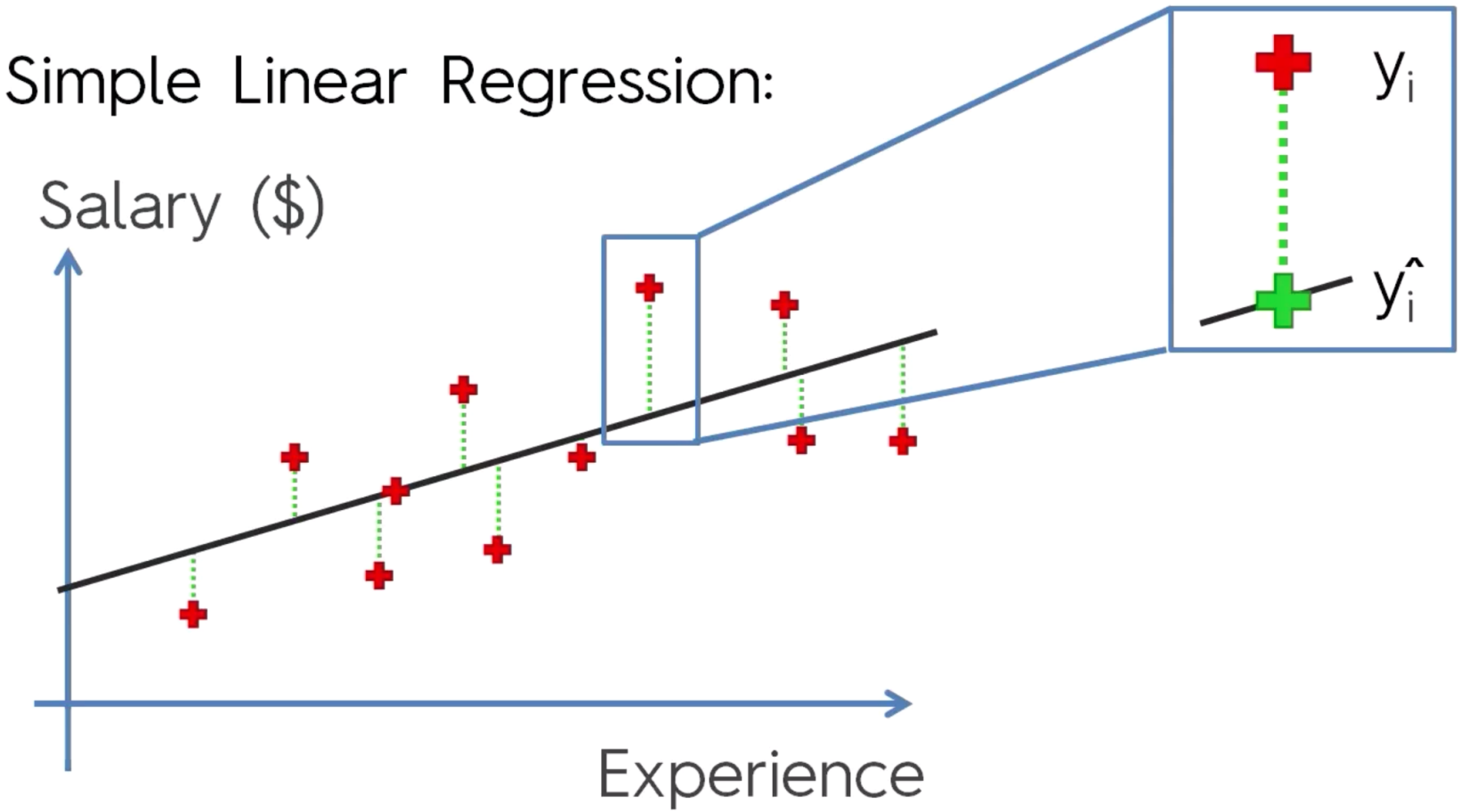
Simple Linear Regression:



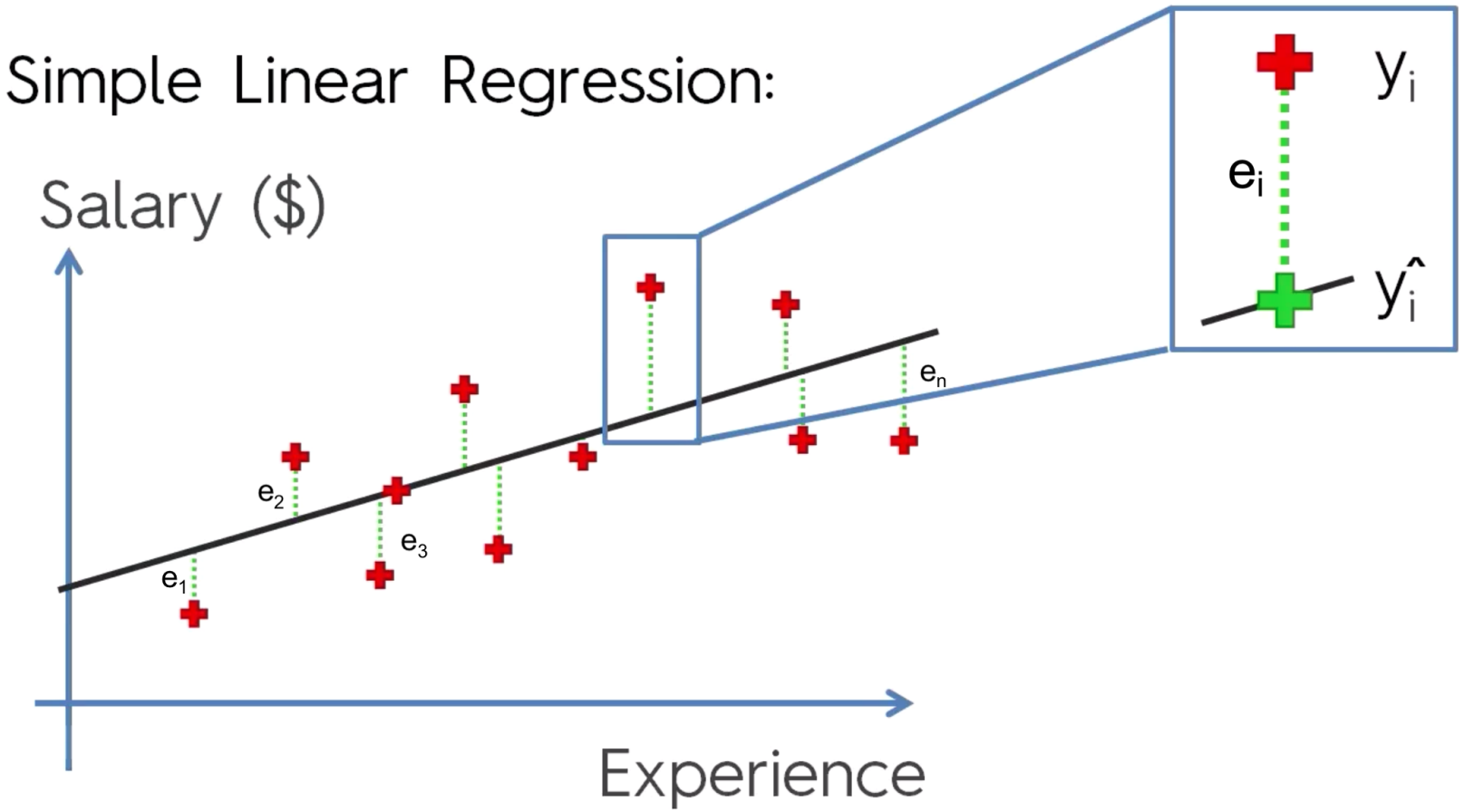
Salary and Work Experience



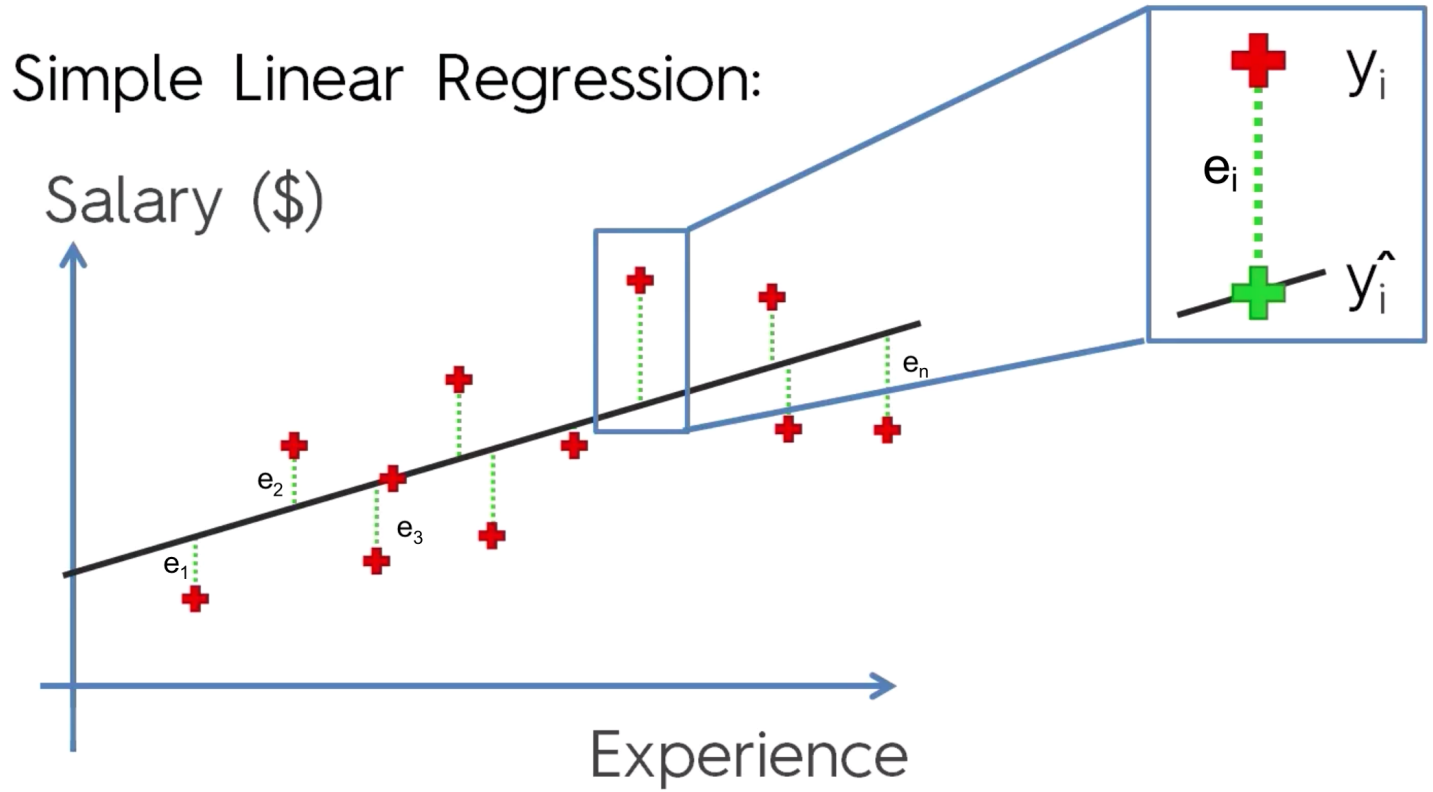
Salary and Work Experience



Salary and Work Experience



Salary and Work Experience

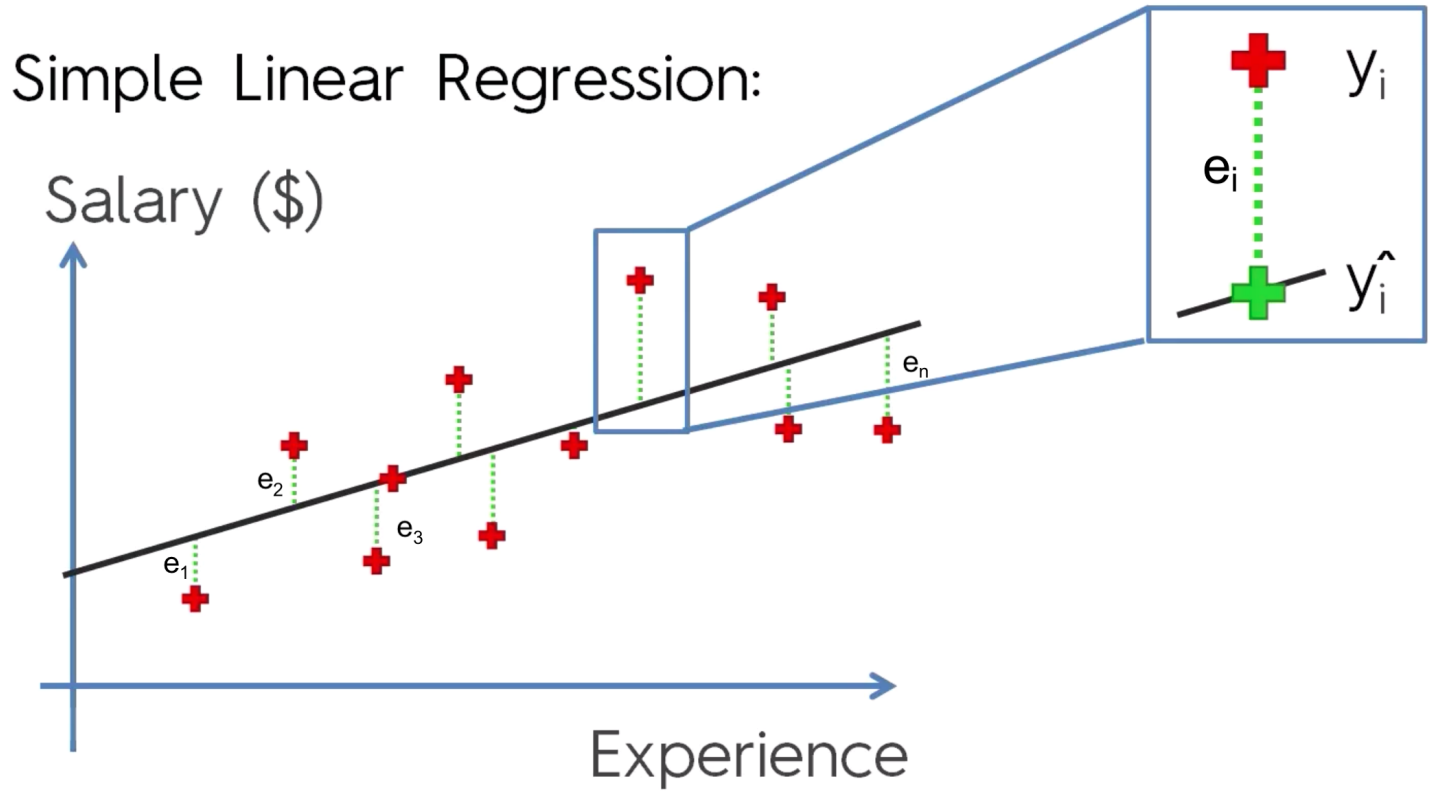


$$\min \rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

RSS - residual sum of squares

$$\min \rightarrow \text{RSS} = \sum_{i=1}^{n} e_n^2$$

Salary and Work Experience

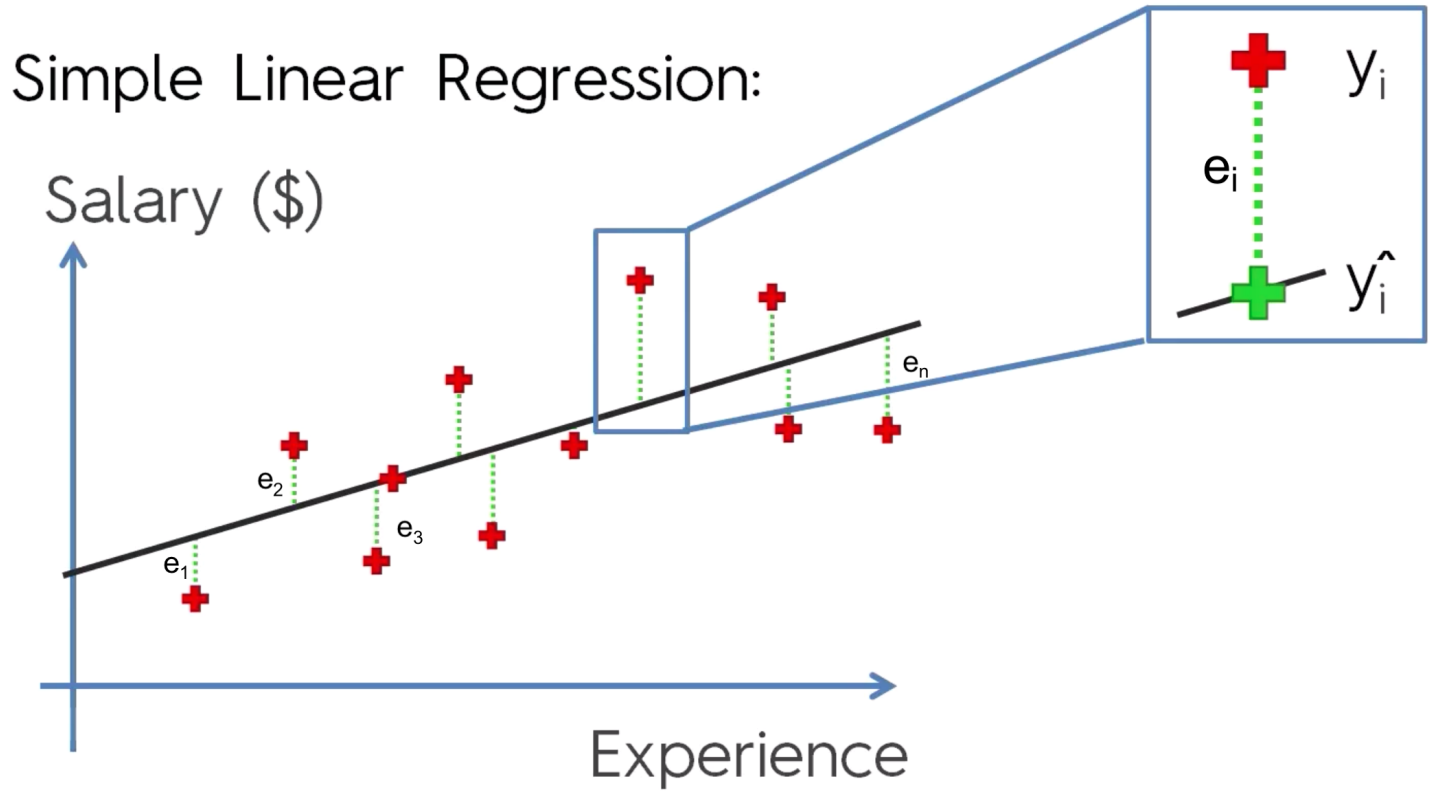


$$\min \rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

RSS - residual sum of squares

$$\min \rightarrow \text{RSS} = \sum_{i=1}^{n} e_n^2$$

Salary and Work Experience



$$\min \rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

RSS - residual sum of squares

$$\min \rightarrow \text{RSS} = \sum_{i=1}^{n} e_n^2$$

Ordinary Least Squares Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Two Components:

- Deterministic: $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$
- Stochastic: ϵ

$$E(Y_i|X_i) + \epsilon = \beta_0 + \beta_1 X_i + \epsilon$$

Stochastic Error Term

At least 3 sources of variation in Y outside that introduced by Xs:

- The underlying theoretical equation might have a different functional form (or shape) than the one chosen for the regression.
- Many minor influences on Y are omitted from the equation (for example, because data are unavailable).
- It is virtually impossible to avoid some sort of measurement error in the dependent variable.

Stochastic Error Term

At least 3 sources of variation in Y outside that introduced by Xs:

- The underlying theoretical equation might have a different functional form (or shape) than the one chosen for the regression.
- Many minor influences on Y are omitted from the equation (for example, because data are unavailable).
- It is virtually impossible to avoid some sort of measurement error in the dependent variable.

Stochastic Error Term

At least 3 sources of variation in Y outside that introduced by Xs:

- The underlying theoretical equation might have a different functional form (or shape) than the one chosen for the regression.
- Many minor influences on Y are omitted from the equation (for example, because data are unavailable).
- It is virtually impossible to avoid some sort of measurement error in the dependent variable.

Stochastic Error Term

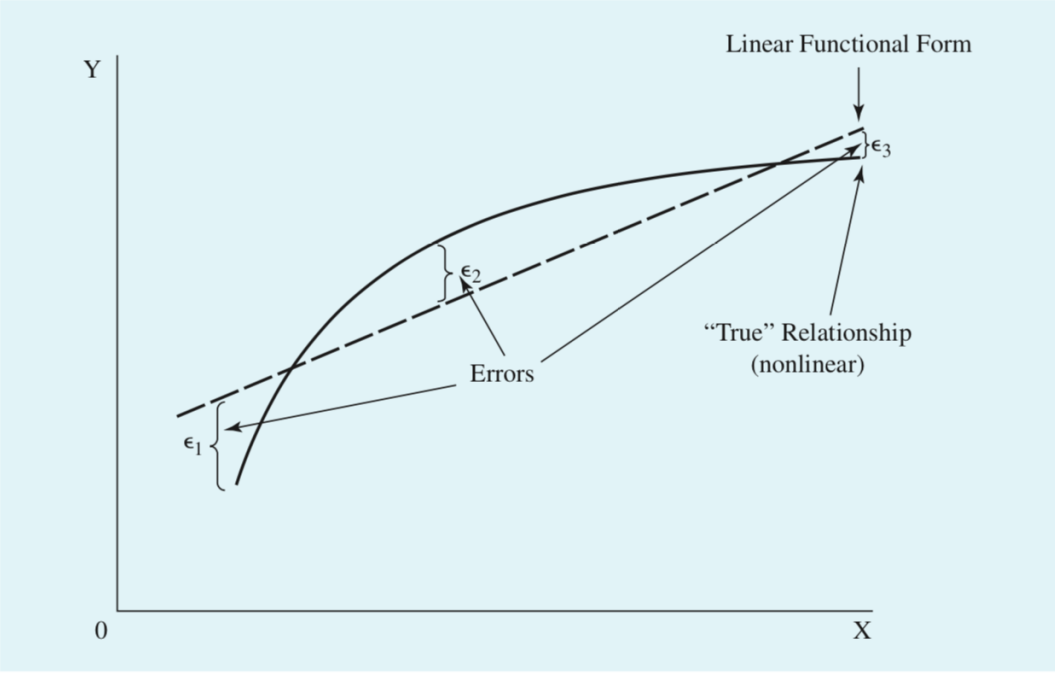
At least 3 sources of variation in Y outside that introduced by Xs:

- The underlying theoretical equation might have a different functional form (or shape) than the one chosen for the regression.
- Many minor influences on Y are omitted from the equation (for example, because data are unavailable).
- It is virtually impossible to avoid some sort of measurement error in the dependent variable.

Stochastic Error Term Example

Aggregate consumption as a function of agg. disposable income:

- The underlying consumption function may be nonlinear, but a linear consumption function might be estimated.

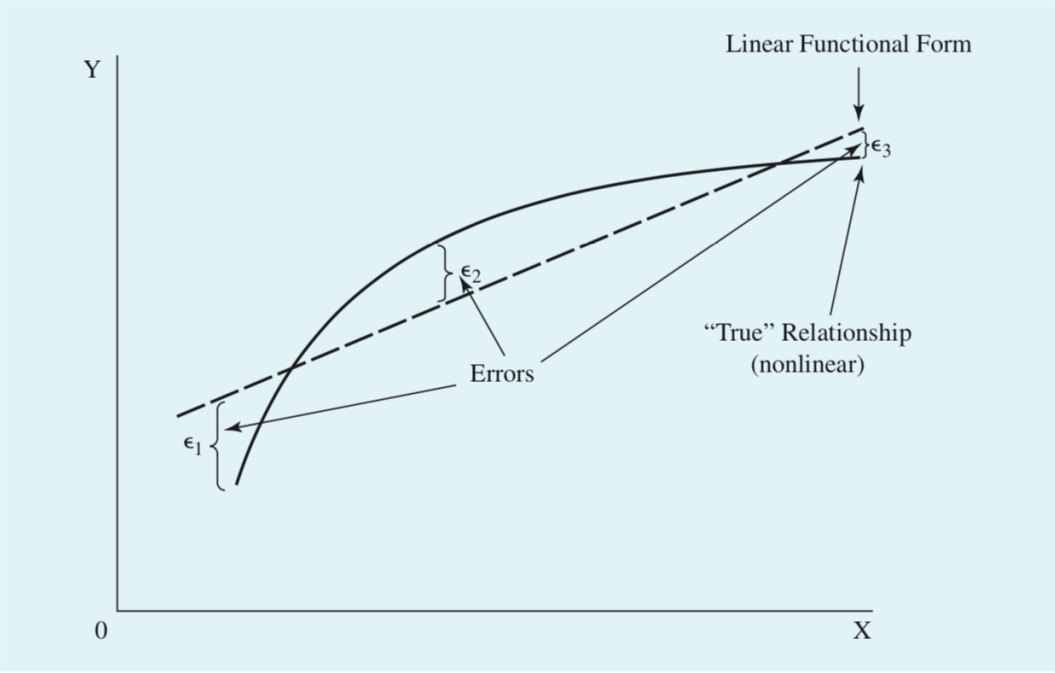


- Consumption may have been less because of uncertainty over the economy. Uncertainty – omitted variable – will likely end up in the stochastic error term.
- Error in the measurement of consumption in the National Income Accounts.

Stochastic Error Term Example

Aggregate consumption as a function of agg. disposable income:

- The underlying consumption function may be nonlinear, but a linear consumption function might be estimated.

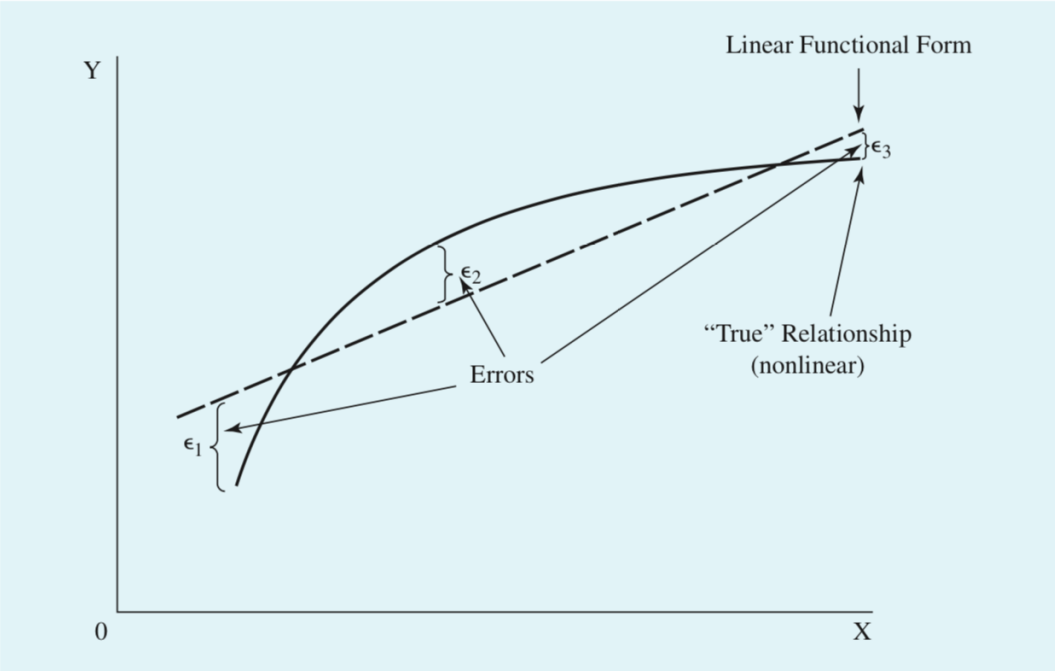


- Consumption may have been less because of uncertainty over the economy. Uncertainty – omitted variable – will likely end up in the stochastic error term.
- Error in the measurement of consumption in the National Income Accounts.

Stochastic Error Term Example

Aggregate consumption as a function of agg. disposable income:

- The underlying consumption function may be nonlinear, but a linear consumption function might be estimated.



- Consumption may have been less because of uncertainty over the economy. Uncertainty – omitted variable – will likely end up in the stochastic error term.
- Error in the measurement of consumption in the National Income Accounts.

Estimated Coefficients

After deciding on the model, we must quantify it:

- Theoretical model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon$
- Estimated Model: $\hat{Y}_i = 9.2 + 0.7 X_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Estimated Coefficients

After deciding on the model, we must quantify it:

- Theoretical model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon$
- Estimated Model: $\hat{Y}_i = 9.2 + 0.7 X_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Estimated Coefficients

After deciding on the model, we must quantify it:

- Theoretical model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon$
- Estimated Model: $\hat{Y}_i = 9.2 + 0.7 X_i$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Estimated Coefficients

After deciding on the model, we must quantify it:

- Theoretical model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon$
- Estimated Model: $\hat{Y}_i = 9.2 + 0.7 X_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

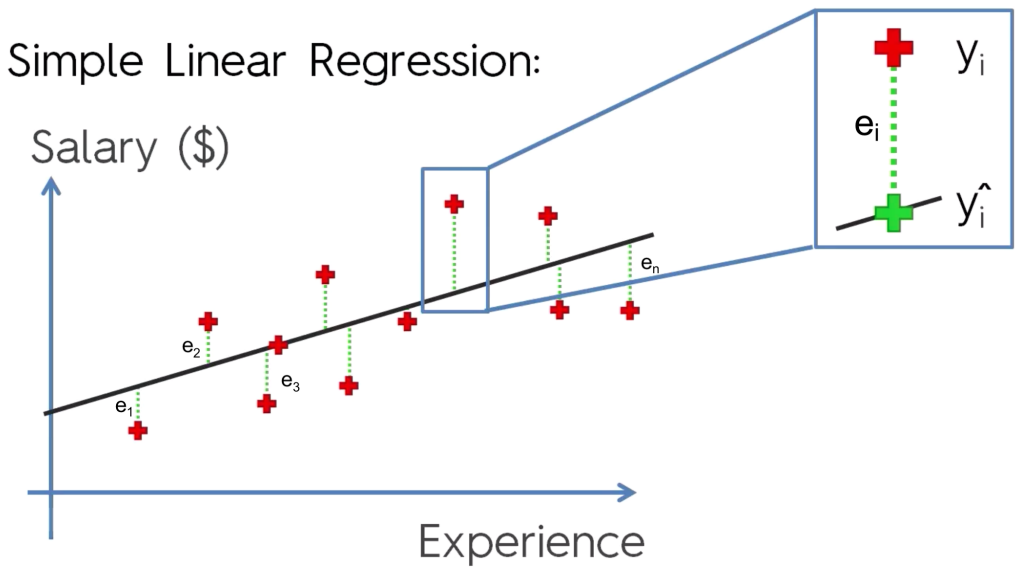
Estimated Coefficients

After deciding on the model, we must quantify it:

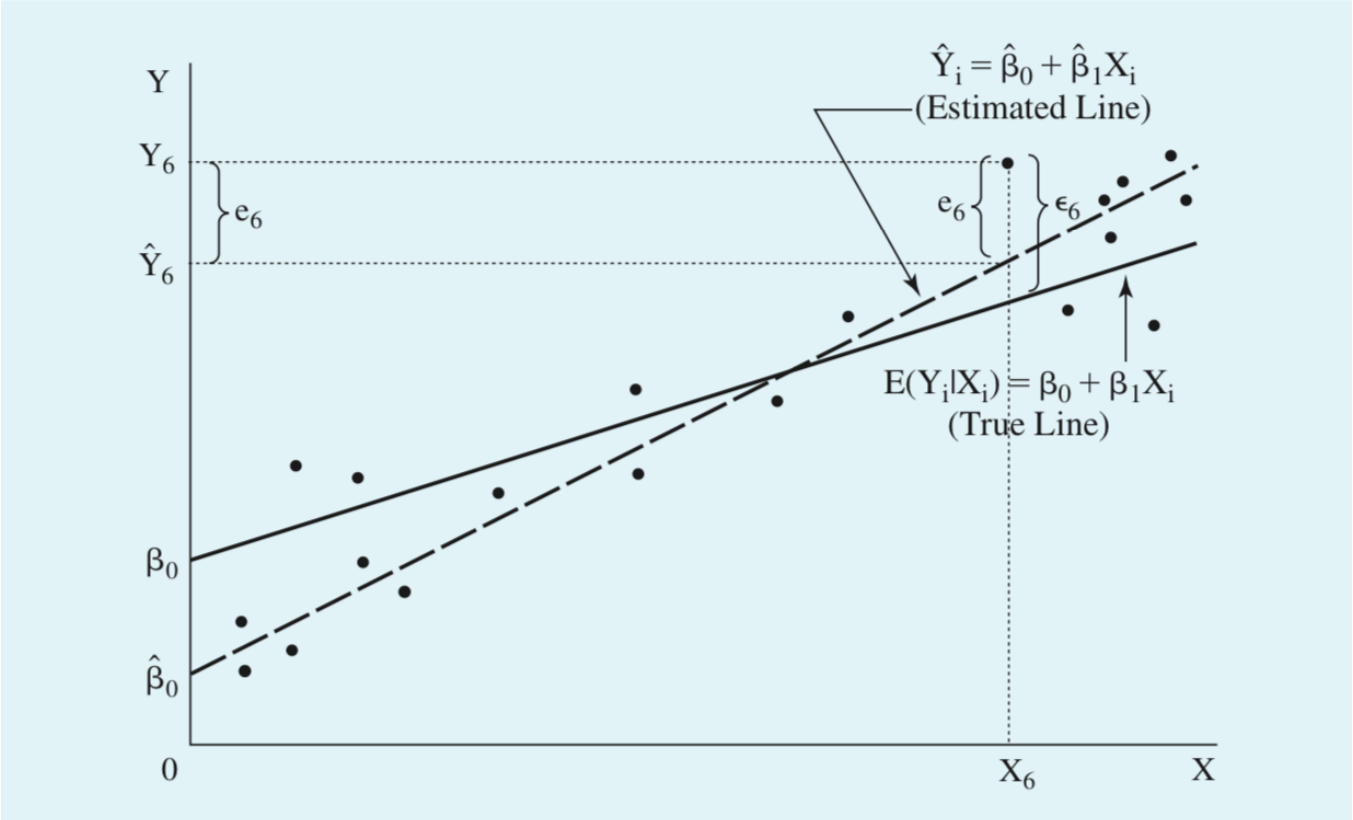
- Theoretical model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon$
- Estimated Model: $\hat{Y}_i = 9.2 + 0.7 X_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$\hat{Y}_i$ is the estimated value of Y_i and as such our prediction of $E(Y_i|X_i)$



Estimated versus True Relationship

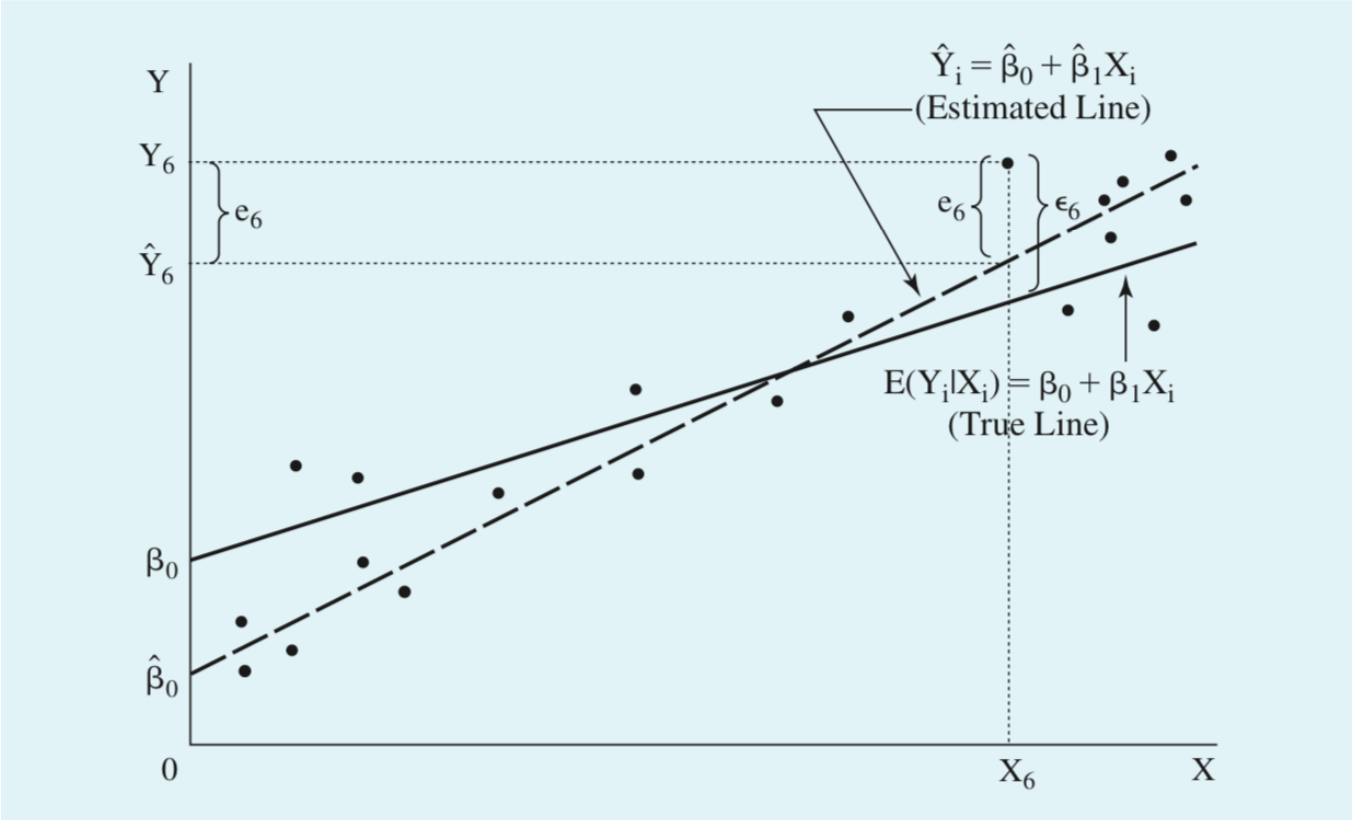


True Regression Equation	Estimated Regression Equation
β_0	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$
ϵ_i	e_i

Residual
$e_i = Y_i - \hat{Y}_i$

Error Term
$\epsilon_i = Y_i - E(Y_i|X_i)$

Estimated versus True Relationship

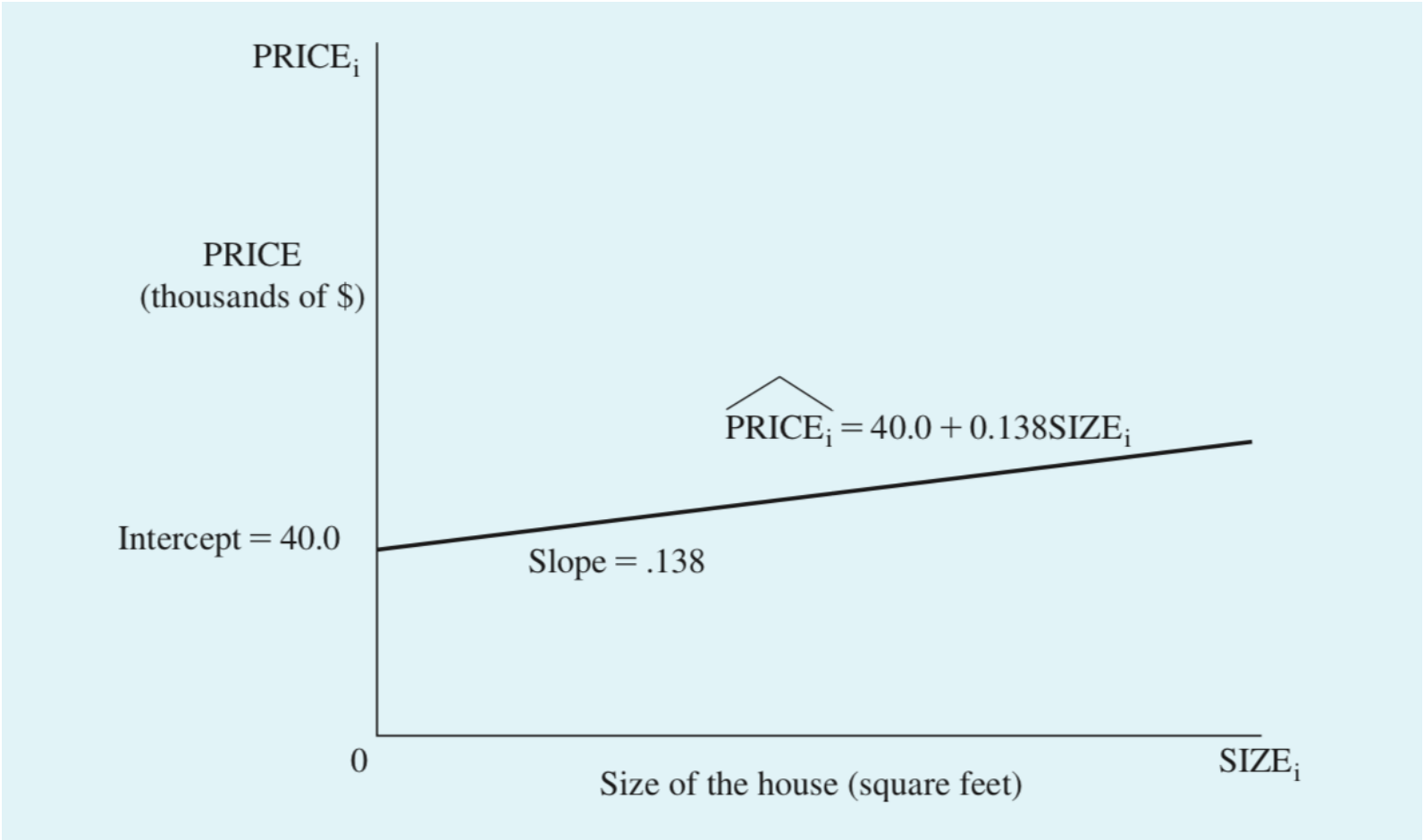


True Regression Equation	Estimated Regression Equation
β_0	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$
ϵ_i	e_i

Residual
$e_i = Y_i - \hat{Y}_i$

Error Term
$\epsilon_i = Y_i - E(Y_i|X_i)$

House Price versus Size



$\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} [\,(X_i - \bar{X})\;(Y_i - \bar{Y})\,]}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\text{min} \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\text{min } f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0, \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$
2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\min \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0,\ \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$
2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2}$$

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\text{min} \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\text{min } f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0,\ \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$
2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2}$$

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\min \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min\ f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0,\ \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2}$$

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\min \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min\ f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0,\ \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$
2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2}$$

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\min \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0,\ \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

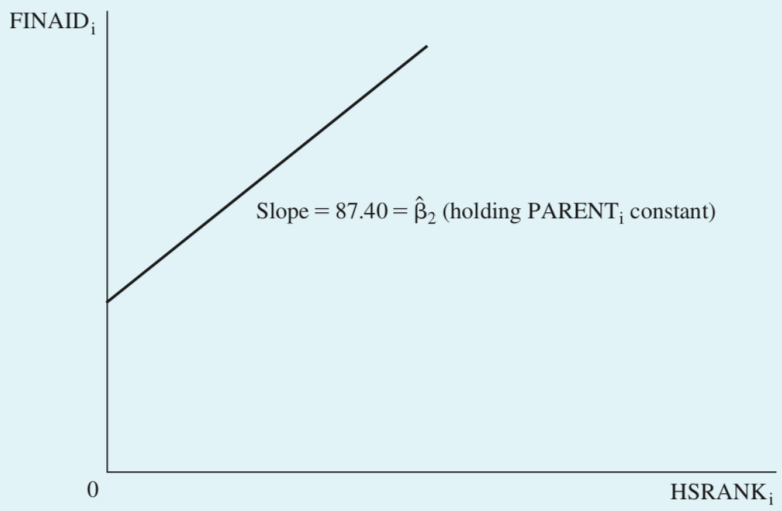
1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$
2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

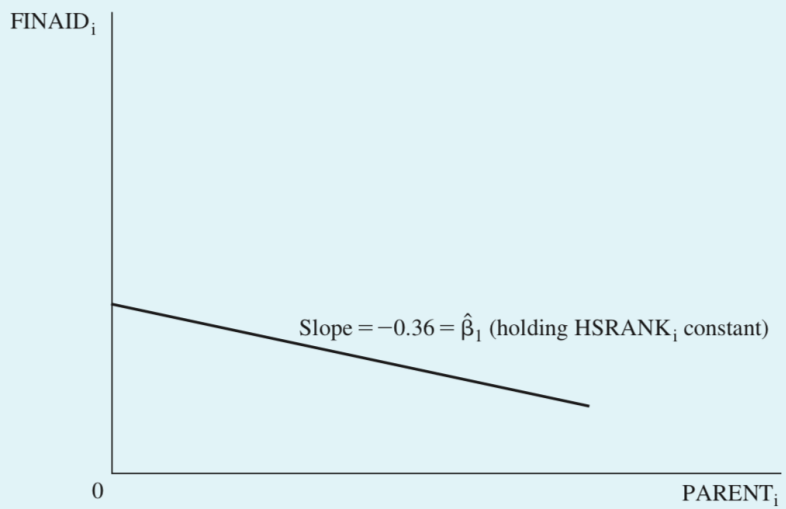
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2}$$

Multivariate OLS

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\underset{-}{} \qquad \underset{+}{}$$
$$\text{FINAID}_i = \beta_0 + \beta_1 \text{PARENT}_i + \beta_2 \text{HSRANK}_i + \epsilon_i$$



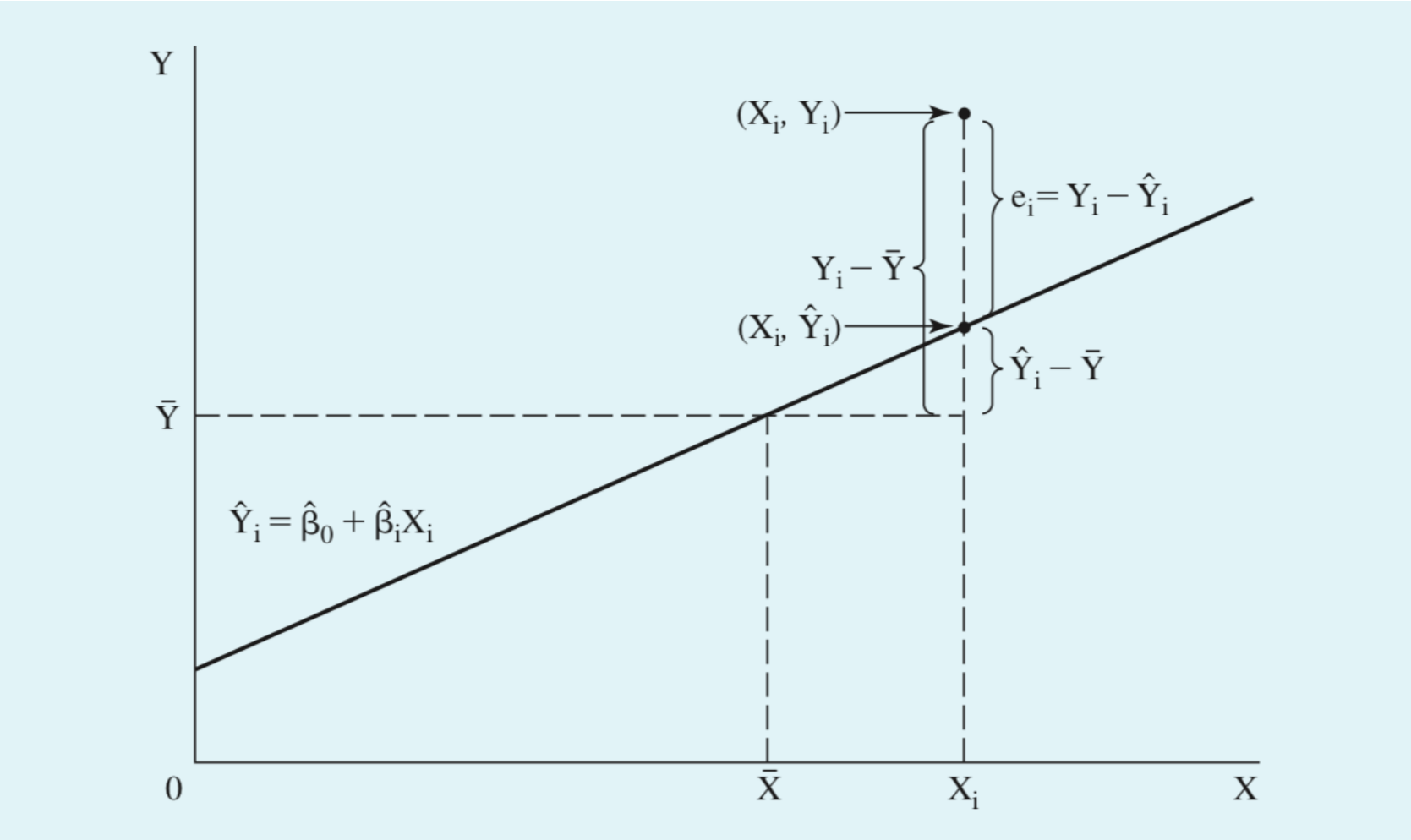


TSS, ESS, RSS

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i e_i^2$$

Total Sum of Squares (TSS)	=	Explained Sum of Squares (ESS)	+	Residual Sum of Squares (RSS)

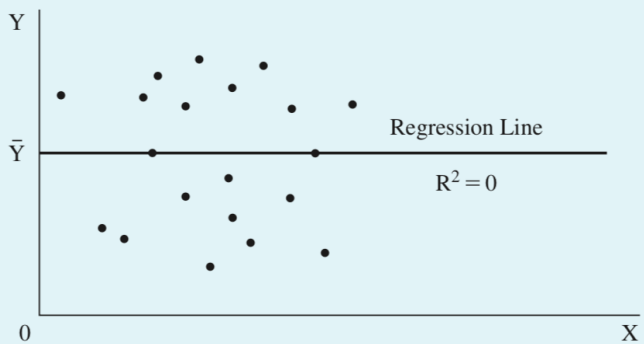
TSS, ESS, RSS and the Regression Line

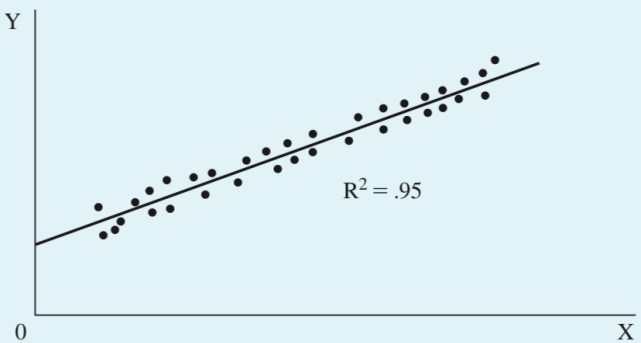


Coefficient of Determination R^2

To judge the goodness of fit of our model we use R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$





Adjusted $\bar{R}^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Problems with R^2

- Adding a variable cannot change TSS, but in most cases the added variable will reduce RSS, so R^2 will rise.
- It also lessens the degrees of freedom, or the excess of the number of observations (N) over the number of coefficients (including the intercept) estimated (K + 1).

Adjusted $\bar{R}^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Problems with R^2

- Adding a variable cannot change TSS, but in most cases the added variable will reduce RSS, so R^2 will rise.
- It also lessens the degrees of freedom, or the excess of the number of observations (N) over the number of coefficients (including the intercept) estimated (K + 1).

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(N - K - 1)}{\sum (Y_i - \bar{Y})^2/(N - 1)}$$